We focus in this paper on high-dimensional regression problems where each regressor can be associated to a location in a physical space, or more generally a generic geometric space. Such problems often employ sparse priors, which promote models using a small subset of regressors. To increase statistical power, the so-called multi-task techniques were proposed, which consist in the simultaneous estimation of several related models. Combined with sparsity assumptions, it lead to models enforcing the active regressors to be shared across models, thanks to, for instance $\ell_1/\ell_q$ norms. We argue in this paper that these techniques fail to leverage the spatial information associated to regressors. Indeed, while sparse priors enforce that only a small subset of variables is used, the assumption that these regressors overlap across all tasks is overly simplistic given the spatial variability observed in real data. In this paper, we propose a convex regularizer for multi-task regression that encodes a more flexible geometry. Our regularizer is based on unbalanced optimal transport (OT) theory, and can take into account a prior geometric knowledge on the regressor variables, without necessarily requiring overlapping supports. We derive an efficient algorithm based on a regularized formulation of OT, which iterates through applications of Sinkhorn’s algorithm along with coordinate descent iterations. The performance of our model is demonstrated on regular grids with both synthetic and real datasets as well as complex triangulated geometries of the cortex with an application in neuroimaging.
to consider that each vector of regression coefficients is sparse, and that a common set of active features is shared across all tasks. This intuition has led to several seminal proposals of Lasso-type models, called multi-task Lasso (MTL) or multi-task feature learning (MTFL) [Argyriou et al., 2007, Obozinski and Taskar, 2006]. Both approaches are based on convex $\ell_1/\ell_2$ group-Lasso norms that promote block sparse solutions.

An issue alluded to by Negahban and Wainwright [2008] is that perfect overlap between all tasks can be a too extreme assumption. To understand how to go beyond this binary idea that active coefficients are the same or not, one can notice that in the context of features mapping to physical locations, employing an $\ell_1/\ell_q$ norm means assuming that exactly the same locations in the physical space, brain or genome are active for each experiment or patient. This is clearly not realistic in several problems [Gramfort et al., 2015].

Our contribution. Our work aims to relax the assumption of perfect overlap across tasks. To do so, we propose to handle non-overlapping supports in standard multi-task models using an optimal transport metric between the parameters of our regression models. Optimal transport (OT) has recently gained considerable popularity in signal processing and machine learning problems. This recent outburst of OT applications can be explained by three factors: the inherent ability of OT theory to compute a meaningful distance between probability measures with non-overlapping supports, faster algorithms to compute that metric using entropic regularization [Cuturi, 2013], and their elegant extension to handle non-normalized measures [Chizat et al., 2017] at no additional computational cost. Our convex formulation exploits these strengths and applies them to a more general setting in which we consider (signed) vectors. In practice, our regularized problem is optimized using alternating updates, namely fast proximal coordinate descent and Sinkhorn’s algorithm. Sinkhorn iterations are matrix-matrix products which can be sped up on parallel platforms such as GPUs. Our experiments on both synthetic and real data show that our OT model outperforms the state of the art by empirical evidence displaying improved performance over block-norm methods. However, experiments show a degraded performance as the overlap between the supports of relevant regressors shrinks. A different approach is proposed by Hernandez-Lobato et al. [2015] where they consider a sparse multi-task regression with outlier tasks and outlier features (non-overlapping features). They introduce a Bayesian model built on a prior distribution with a set of binary latent variables for each feature and each task. Han and Zhang [2015] propose to learn a tree structure on the features, with inner nodes defined as spatially pooled features. The main advantage of this approach is that no assumptions are made on how tasks are related (as in the work of Jawanpuria and Jagarlapudi [2012]). However, the inner nodes will be selected if the supports across tasks do not overlap, resulting in spatially smeared coefficients. Learning how tasks are related adaptively is a potential extension of our work.

This paper is organized as follows. Section 2 introduces our main contribution, the multi-task Wasserstein (MTW) model. We present in Section 3 a computationally efficient optimization strategy to tackle the MTW inference problem. Section 4 demonstrates with multiple experiments the practical benefits of our model compared to Lasso-type models.

Notation. We denote by $I_p$ the vector of ones in $\mathbb{R}^p$. Given an integer $d \in \mathbb{N}$, $[d]$ stands for $\{1, \ldots, d\}$. The set of vectors in $\mathbb{R}^p$ with non-negative (resp. positive) entries is denoted by $\mathbb{R}^p_+$ (resp. $\mathbb{R}^p_{++}$). On matrices, log, exp and the division operator are applied element-wise. We use $\odot$ for the element-wise multiplication between matrices or vectors. If $X$ is a matrix, $X_i$ denotes its $i$th row and $X_{ij}$ its $j$th column. We define the Kullback-Leibler (KL) divergence between two positive vectors by $\text{KL}(x, y) = \langle x, \log(x/y) \rangle + \langle y-x, I_p \rangle$ with the continuous extensions $0 \log(0/0) = 0$ and $0 \log(0) = 0$. We also use the convention that for $x \neq 0$, $\text{KL}(x|0) = +\infty$. The entropy of $x \in \mathbb{R}^n$ is defined as $E(x) = -\langle x, \log(x) - I_p \rangle$. Finally, for any vector $u \in \mathbb{R}^p$, the support of $u$ is $\mathcal{S}_u = \{i \in [p], u_i \neq 0\}$.

2 Multi-task Wasserstein model

Multi-task regression. Consider $T$ datasets of labeled vectors $(X^t, Y^t) \in \mathbb{R}^{n_t \times p} \times \mathbb{R}^{n_t}$, where $n_t$ is the sample size of each set, and $p$ is the dimension of the common space in which all observations lie. Our aim is to estimate, in a high-dimensional regime $n_t \ll p$, $T$ linear regression models:

$$Y^t = X^t \theta^t + \epsilon^t, \quad t \in [T],$$

where $\theta^1, \ldots, \theta^T \in \mathbb{R}^p$ are regression coefficients to be estimated from the samples $X^t$ with associated
labels $Y^t$, and $\epsilon_1, \ldots, \epsilon_T \in \mathbb{R}^n$ are additive noise terms assumed to be i.i.d centered Gaussian variables with the same variance $\sigma^2 I_n$. For simplicity, we will assume from now on that $n_t = n$.

**Multi-task consensus through Geometric Variance.** The idea behind multi-task learning is to estimate $\theta^1, \ldots, \theta^T$ jointly, using a regularization term $J$ that promotes some form of similarity between them. All multi-task regression models can then be written:

$$
\min_{\theta^1, \ldots, \theta^T} 1 \over 2n \sum_{t=1}^T \|X^t \theta^t - Y^t\|^2 + J(\theta^1, \ldots, \theta^T). \tag{1}
$$

We propose to use a regularizer that promotes not only sparse solutions, but also some form of “geometric” consensus across all $\theta^1, \ldots, \theta^T$ through the use of an arbitrary discrepancy function $\Delta : \mathbb{R}^p \times \mathbb{R}^p \rightarrow \mathbb{R}$, writing $J(\theta^1, \ldots, \theta^T) \overset{\text{def}}{=} \min_{\theta \in \mathbb{R}^p} H(\theta^1, \ldots, \theta^T; \theta)$, where for regularization parameters $\mu \geq 0$ and $\lambda > 0$,

$$
H(\theta^1, \ldots, \theta^T; \theta) \overset{\text{def}}{=} \frac{\mu}{T} \sum_{t=1}^T \Delta(\theta^t, \theta) + \frac{\lambda}{T} \sum_{t=1}^T \|\theta^t\|_1, \tag{2}
$$

We call the first quantity a geometric variance because it boils down to the usual variance when $\Delta$ is the squared Euclidean distance. Indeed, the minimization of $\theta$ in $J$ would return the mean of all $\theta^t$, and the first sum in $H$ would then be the variance of these vectors.

**An OT Discrepancy for Vectors in $\mathbb{R}^p$.** To quantify the geometric variance, we propose to use a new generalized OT metric, that can leverage the fundamental ability of Wasserstein distances to provide a meaningful meta-distance between vectors when a metric on the bins of these vectors is known. However, since OT metrics are defined for positive and normalized vectors, using them in our setting requires some adaptation. Similarly to [Profeta and Sturm, 2018, Mainini, 2012], we propose to split each vector in its positive and negative parts. More formally we write $(x_+, x_-) \in \mathbb{R}^p_+$ such that $x = x_+ + x_-$ by setting $x_+ = \max(x, 0)$ applied elementwise. Next, denoting $W$ the unbalanced Wasserstein distance introduced by Chizat et al. [2017] and described in detail in the next paragraph, we consider in the rest of this work for two arbitrary vectors $x, y \in \mathbb{R}^p$:

$$
\Delta(x, y) \overset{\text{def}}{=} W(x_+, y_+) + W(x_-, y_-). \tag{3}
$$

When $\mu = 0$, (2) boils down to the penalty of $T$ independent Lasso models, one for each task. When the $\theta^t$ are fixed, the minimization w.r.t. $\theta_+$ (resp. $\theta_-$) consists in estimating the barycenter of the $\theta^t_+$ (resp. $\theta^t_-$) according to the metric $W$. When $\lambda = 0$, one forces all the coefficients to be closer according to $W$.

**Unbalanced Wasserstein distance $W$.** The reason why optimal transport distances fit our framework is that they can leverage knowledge on the geometry of regressors, in situations such as those presented in the introduction. In OT, that knowledge is known as a ground metric. When working in $\mathbb{R}^p$, this ground metric can be seen as a substitution cost matrix between all $p$ regressors, and is given as a matrix $M \in \mathbb{R}^{p \times p}_+$ of pairwise distances between bins. Following the historical analogy of mass displacement cost, $M_{ij}$ represents the cost to move one unit of mass from location $i$ to location $j$. In the current context, $M$ may come from the knowledge that features map to certain spatial positions. For instance, in applications where features correspond to positions $(x_1, \ldots, x_p)$ in a Euclidean space, the standard cost matrix is given by $M_{ij} = \|x_i - x_j\|^2$.

As proposed in [Frogner et al., 2015, Chizat et al., 2017], an optimal transport cost between two nonnegative vectors $\theta_1$ and $\theta_2$ can be defined by seeking a transport plan $P \in \mathbb{R}^{p \times p}_+$ that: (i) achieves low transport cost $\langle P, M \rangle$; (ii) has marginals $P1$ (resp. $P^T1$) that are as close as possible to $\theta_1$ (resp. $\theta_2$) in KL sense and (iii) has high entropy. These three requirements are reflected in the definition:

$$
W(\theta_1, \theta_2) \overset{\text{def}}{=} \min_{P \in \mathbb{R}^{p \times p}_+} G(P, \theta_1, \theta_2), \tag{4}
$$

where

$$
G(P, \theta_1, \theta_2) = (\langle P, M \rangle - \varepsilon E(P) + \gamma KL(P1|\theta_1) + \gamma KL(P^T1|\theta_2), \tag{5}
$$

and $\varepsilon, \gamma > 0$ are parameters providing a tradeoff between these different objectives.

Large values of $\gamma > 0$ tend to strongly penalize unbalanced transports, and as a result penalize discrepancies between the marginals of $P$ and $\theta_1, \theta_2$. The entropy regularization, first introduced by Cuturi [2013], makes the problem strictly convex and computationally faster to solve. A crucial feature of this definition is that the resolution of (4) does not require computing nor storing in memory any optimal plan $P^*$. Instead, one can study its Fenchel-Rockafellar dual problem given by:

$$
W(\theta_1, \theta_2) = \max_{u, v \in \mathbb{R}^p} \left\{ -\varepsilon \langle u \otimes v - 1, K \rangle - \gamma \langle u^{-\frac{p}{\gamma}} - 1, \theta_1 \rangle - \gamma \langle v^{-\frac{p}{\gamma}} - 1, \theta_2 \rangle \right\}, \tag{6}
$$

Performing alternating gradient ascent on (6) amounts to computing matrix scalings of a generalized Sinkhorn algorithm (see Section 3).
Well-posedness. We show in this paragraph that a minimizer of (1) exists. To do so, we must prove that the objective function is continuous and coercive.

Lemma 1. For any \( \theta_1, \theta_2 \in \mathbb{R}_+^p \)

\[
W(\theta_1, 0) = W(0, \theta_2) = W(0, 0) = 0
\]

Proof. We show that \( W(0, \theta_2) = 0 \). The result follows directly from the definition of the KL divergence. Let \( P \in \mathbb{R}_+^{p \times p} \). We have KL\((P1, 0) = 0 \) if \( P = 0 \) and \(+\infty\) otherwise. Thus, the minimizer of \( G(P, 0, \theta_2) = P^* = 0 \) and we have \( W(0, \theta_2) = G(0, \theta_1, 0) = 0 \). The same reasoning applies to prove \( W(\theta_1, 0) = W(0, 0) = 0 \).

Proposition 1. The extension (3) preserves the continuity of \( H \) at 0.

Proof. Since \( H \) is separable across the \( (\theta_i) \), we only need to prove that for \( \theta, \bar{\theta} \in \mathbb{R}_+^p \)

\[
\lim_{(\theta, \bar{\theta}) \to (0, 0)} W(\theta, \bar{\theta}) = \lim_{\theta, \bar{\theta} \to 0} W(\theta, \bar{\theta}) = \lim_{\theta, \bar{\theta} \to 0} W(\theta, 0) = 0 \quad \text{if} \quad i, j \in [p] \quad \text{Suppose} \quad \theta_i, \bar{\theta}_j \neq 0. \quad \text{G is smooth, convex and coercive w.r.t \( P \).}
\]

The first order optimality condition reads:

\[
M + \varepsilon \log(P_{ij}) + \gamma \log((P1)_{ij}(P^\top 1)_{ij}) = \gamma \log(\bar{\theta}_i \bar{\theta}_j)
\]

\[\Leftrightarrow \exp(M)\gamma(P_{ij}) = ((P1)_{ij}(P^\top 1)_{ij}) \gamma = (\bar{\theta}_i \bar{\theta}_j) \gamma
\]

When \( (\theta, \bar{\theta}) \rightarrow (0, 0) \) \( (P_{ij})^\varepsilon ((P1)_{ij}(P^\top 1)_{ij})^\gamma \rightarrow 0 \) and \( P \) is non-negative we have \( \forall i, j, P_{ij} \rightarrow 0 \), \( i.e \) \( P \rightarrow 0 \). The continuity of \( G \) with respect to \( P \)
leads to \( \lim_{(\theta, \bar{\theta}) \to (0, 0)} W(\theta, \bar{\theta}) = 0 \). Lemma 1 guarantees \( \lim_{(\theta, \bar{\theta}) \to (0, 0)} W(\theta, 0) = \lim_{\theta, \bar{\theta} \to 0} W(\theta, 0) = 0 \).

Proposition 1 shows that our extension still guarantees that \( H \) is continuous at 0. Now we show that the loss function in (1) is coercive.

Proposition 2. The loss function in (1) is coercive.

Proof. Let’s prove that \( W \) is bounded from below. Since KL is non-negative, and \( (P, M) \geq 0 \), we have \( W(\theta_1, \theta_2) \geq \min_{P \in \mathbb{R}_+^{p \times p}} -E(P) \) which is minimized at \( P^*_ij = 1 \forall i, j \). Thus \( W(\theta_1, \theta_2) \geq -\varepsilon P^2 \). Thus, given that the \( \ell_1 \) norm is non-negative, \( H \) is also bounded from below. The coercivity of the loss function follows from the coercivity of the quadratic loss.

3 Efficient Optimization of MTW

Loss function. We solve MTW by alternating minimization on the positive and negative parts of the regression coefficients \( \theta \equiv (\theta_1, \ldots, \theta^T) \) and those of \( \bar{\theta} \).

We will use what follows bold symbols for sequences of the form \( \mathbf{z}^T = (z^1, \ldots, z^T) \). Let \( P_1 \) and \( P_2 \) denote respectively the optimal transport plans linking \( \theta \) with \( \bar{\theta}_+ \) and \( \theta_- \) with \( \bar{\theta}_- \), and \( m_1 \) and \( m_2 \) their respective left marginals. Combining (1), (2) and (4), the cost function to minimize is given by:

\[
L(\theta; P_1; P_2; \bar{\theta}) = \sum_{t=1}^T \left[ \frac{1}{2\mu} \left< X^t \theta^t - Y^t, X^t \theta^t - Y^t \right> + \lambda \| \theta^t \|_1 \right] + \frac{\mu}{T} \left( G(P_1^t; \theta^t_+, \bar{\theta}_+) + G(P_2^t; \theta^t_-, \bar{\theta}_-) \right)
\]

\[L \] is jointly convex in all its variables (since the Kullback-Leibler is jointly convex, proof in Supplementary materials) and the remaining terms are convex and not coupled. The straightforward solution is to minimize \( L \) by block coordinate descent. Since the minimization with respect to variables \( (P_1^t; \theta^t_+, \bar{\theta}_+) \) and \( (P_2^t; \theta^t_-, \bar{\theta}_-) \) is similar, we only detail hereafter the minimization with respect to \( (P_1^t; \theta^t_+, \bar{\theta}_+) \). The full optimization strategy is provided in Algorithm 1. We alternate with respect to \( (P_1, \bar{\theta}_+) \) and each \( \theta^t_+ \), which can be updated independently and therefore in parallel. We now detail the two steps of the procedure.

Barycenter update. For fixed \( \theta_+ \), minimizing with respect to \( (P_1, \bar{\theta}_+) \) boils down to the unbalanced Wasserstein barycenter computation of [Chizat et al., 2017] which generalizes previous work by Agueh and Carlier [2011] to compute the minimizer of \( \min_{\theta_+ \in \mathbb{R}_+^p} \frac{1}{T} \sum_{t=1}^T W(\theta^t_+, \bar{\theta}_+) \). This is equivalent to minimizing simultaneously in \( P_1^t, \ldots, P_t_1 \in \mathbb{R}_+^{p \times p} \) and \( \theta_+ \in \mathbb{R}_+^p \) the objective:

\[
\sum_{t=1}^T \KL(P_1^t; K) + \sum_{t=1}^T \KL(P_1^t; \theta^t_+) + \sum_{t=1}^T \KL(P_1^t; \bar{\theta}_+)
\]

As pointed out by Chizat et al. [2017] and recalled in (6), Fenchel-Rockafellar duality allows to minimize over dual variables \( u^t, v^t \in \mathbb{R}^p \) instead of considering plans \( P_1^t \in \mathbb{R}_+^{p \times p} \). \( P_1^t \) can be recovered as \( (u^t_i K_{ij} v^t_j)_{ij} \)

and its left marginal, needed for the coefficient update, is given by \( m_1^t \equiv P_1^t 1 = u^t \odot K v^t \). These steps are summarized in Alg. 2. We monitor the largest relative change of barycenter the \( \theta_+ \) to stop our loop.

Coefficients update. Minimizing with respect to one \( \theta^t_+ \) while keeping all other variables fixed to their current estimate yields problem (9), where the \( \ell_1 \) penalty becomes linear due to the positivity constraint.

Given the left marginal \( m_1 \), the problem reads for all \( \theta^t_+ \) (omitting index \( t \)):

\[
\min_{\theta_+ \in \mathbb{R}_+^p} \left[ \frac{1}{2\mu} \left< X \theta_+ - X \bar{\theta}_- - Y, X \theta_+ - X \bar{\theta}_- - Y \right> + \sum_{i=1}^p \frac{\mu_i}{T} (\theta^t_{i+} - m_{i+} \log(\theta^t_{i+}) + \lambda \theta^t_{i+}) \right]
\]
The penalty is a separable sum of convex functions with tractable proximal operators, and therefore (9) can be solved by proximal coordinate descent [Tseng, 2001, Fercoq and Richtárik, 2015]. The following proposition, proved in the appendix, gives a closed-from solution for that proximal operator.

**Proposition 3.** Let $a, b, \alpha \in \mathbb{R}_{++}$. Function $g : x \mapsto (x - a \log(x)) + bx$ is convex on $\mathbb{R}_{++}$, and one has:

$$\text{prox}_g(y) = \frac{1}{2} \left[ -\alpha(b + 1) + y + \sqrt{(\alpha(b + 1) - y)^2 + 4\alpha y} \right].$$

**Entropy regularization.** While large values of $\varepsilon$ (strong entropy regularization) induce undesired blurring, low values tend to cause a well-documented numerical instability [Chizat et al., 2017, Schmitzer, 2016], which can be avoided by moving to the log-domain [Schmitzer, 2016]. Also for experiments performed on regular grids such as images, one should leverage the separability of the kernel $K$ as proposed in [Solomon et al., 2015] to recover far more efficient implementations. This also applies to log-domain computations [Schmitz et al., 2017]. We use in this work these crucial improvements over naive implementations of Sinkhorn algorithms.

**Accelerating convergence with warm-start.** To speed up convergence, we initialize the Sinkhorn scaling vectors to their previous values, kept in memory between two barycenter computations. This does not affect convergence because of the convexity of the objective function. Note that transport plans $P^1, \ldots, P^T$ are never instantiated, as this would be too costly. We only compute their left marginals $m^1, \ldots, m^T$, which are involved in the coefficients update. We track both the relative evolution of the objective function and that of the norm of the coefficients to terminate the algorithm. Performing less Sinkhorn iterations per barycenter update yields in practice faster convergence, while reaching the same final tolerance threshold. See supplementary materials for an illustration of this trade-off.

**Hyperparameter tuning.** The MTW model has four hyperparameters: $\varepsilon, \gamma, \mu, \lambda$. We provide in this section practical guidelines to set parameters $\varepsilon$ and $\gamma$ within the unbalanced Wasserstein distance.

**Setting $\varepsilon$.** As mentioned above, entropy regularization speeds up computations but induces blurring. In our experiments we observe that a value of $1/sp$, where $s$ is the median of the ground metric $M$, provides an excellent tradeoff between speed and performance.

**Setting $\gamma$.** In the barycenter definition (8), $\gamma$ controls the influence of the marginals: as $\gamma$ goes to 0, $P$ tends to $K$ since we can ignore marginal constraints. This transport plan, however, only leads to a local blur with no transport, so that the mass of the barycenter $\bar{\theta}1 \rightarrow 0$. To avoid this degenerate behavior, consider the case where $\gamma \gg 1$, so entropy regularization can be neglected in (8). The corresponding approximate objective function is given by:

$$\sum_{t=1}^{T} \left[ \langle P^t, M \rangle + \gamma \text{KL}(P^t|a^t) + \gamma \text{KL}(P^t|\mathbb{1}|a) \right].$$

Deriving the first order conditions, for any $t \in \{T\}$:

$$M_{ij} + \gamma \log \left( \frac{P_{ij}^t}{a_{ij}^t} \right) = 0 \quad \text{and} \quad \bar{a} = \frac{1}{T} \sum_{t=1}^{T} P_{ij}^t \mathbb{1},$$

By combining the two, we get for any $\tau \in [0, 1]$:

$$\gamma \geq \frac{\max M}{\log \tau} \Rightarrow \bar{v} \geq \tau \left( \frac{1}{T} \sum_{t=1}^{T} \sqrt{\psi_t} \right)^2,$$

where $\bar{v}, \psi_1, \ldots, \psi_T$ denote the respective masses of
\[ \tilde{a}, \tilde{a}_1, \ldots, \tilde{a}_T \text{ i.e. } \psi^t = a^t \top \mathbb{1}. \] Therefore, (11) provides an adaptive parametrization of \( \gamma \) that guarantees a lower bound on the mass of \( \tilde{\theta} \) as a fraction of the \( \ell_{0.5} \) pseudo-norm of those of the inputs. In practice, in all experiments we use \( \tau = 0.5 \) and set \( \gamma = \tau \left( \frac{\sum_{t=1}^T \sqrt{\psi_t}}{T} \right)^2 \). Code in Python can be found in https://github.com/hichamjanati/mtw.

With \( \varepsilon \) and \( \gamma \) fixed, only two hyperparameters \( (\mu, \lambda) \) remain. These control respectively the similarity between tasks and sparsity. Setting two parameters is not more than what is required by Dirty models [Jalali et al., 2010] or an Elastic-Net.

4 Experiments

**Benchmarks.** To quantify the benefit of multi-task inference, we use a Lasso estimator independently run on each task as a standard baseline. We compare the performance of our algorithm against Dirty models [Jalali et al., 2010] and Multi-level Lasso [Lozano and Swirszcz, 2012]. When the ground truth is available, we evaluate support identification using the area under the curve (AUC) of precision-recall.

The Group Lasso learning model [Argyriou et al., 2007, Obozinski and Taskar, 2006] (a.k.a MTFL) can be expressed by setting the penalty \( J \) to be an \( \ell_1/\ell_2 \) mixed norm: \( \| \theta \|_{21} = \sum_{j=1}^{p} \sqrt{T} \sum_{t=1}^{T} (\theta^j_t)^2 \). Such a regularization forces all the \( \theta^j \) to have the exact same support, \( S_{\theta^j} = S_{\theta^{j'}} \) for all \( t, t' \). To mance this very strong assumption, Dirty models [Jalali et al., 2010] propose to decompose \( \theta^j = \theta^j_{c} + \theta^j_{e} \), where \( \theta^j_{c} \) is common between all tasks (i.e. \( S_{\theta^j} = S_{\theta^{j'}} \forall t, t' \)) and \( \theta^j_{e} \) is specific to each one. The regularization then writes:

\[
J_{\text{Dirty}}(\theta) = \mu \| (\theta^1_{c}, \ldots, \theta^T_{c}) \|_{21} + \lambda \sum_{t=1}^{T} \| \theta^t_{e} \|_1. \tag{12}
\]

When \( \theta_{c} = 0 \) (resp. \( \theta_{e} = 0 \)) one falls back to a Group Lasso (resp. independent Lasso) estimator [Argyriou et al., 2007, Obozinski and Taskar, 2006].

Multi-level Lasso (MLL) applies instead the \( \ell_1 \) penalty on two levels of a product decomposition \( \theta^j = \bar{C}_j \bar{S}^j \), where \( \bar{C} \in \mathbb{R}^p \) is common across tasks and \( \bar{S}^j \in \mathbb{R}^p \) is task specific. For the model to be identifiable, \( \bar{C} \) is constrained to be non-negative. The (MLL) penalty:

\[
J_{\text{MLL}}(S^1, \ldots, S^T; C) = \mu \| C \|_1 + \lambda \sum_{t=1}^{T} \| S^t \|_1. \tag{13}
\]

As shown by Lozano and Swirszcz [2012], (13) is equivalent to a standard multi-task regression problem with the non-convex regularization:

\[
J(\theta) = \frac{1}{T} \sum_{j=1}^{p} \left( \sum_{t=1}^{T} |\theta^j_t| \right). \tag{14}
\]

4.1 Synthetic data

We simulate 3 coefficients \( (\theta^j)^{1} = 3 \) defined on a 2D grid of shape \((24 \times 24)\), and that each vector of coefficients is 4-sparse: each has only 4 non-zero values (see Figure 1). Each coefficient can be seen as a 24 \( \times \) 24 image. We thus have 3 tasks with \( p = 576 \). The design matrix is obtained by applying a Gaussian filter to the image with standard deviation of 1 pixel, and downsampling the blurred image by taking the mean over \((4 \times 4)\) blocks. This leads to \( n = 36 \) samples. We set the Gaussian noise variance \( \sigma^2 \) so that the signal-noise-ratio (SNR) is equal to 3, with \( \text{SNR}^2 \triangleq \sum_{j} ||X^j \theta^j ||_2^2 / (T \sigma^2) \).

To control the overlap ratio between the supports and guarantee their proximity, we first start by selecting two random pixels and randomly translating the non-overlapping features by a one or two pixels for the corresponding tasks. The coefficient values are drawn uniformly between 20 and 30.

Here coefficients map to image pixels, so we employed the MTW with a non-negativity constraint \( (\theta^j \in \mathbb{R}_+) \). Figure 2 shows the distribution of the best AUC scores of 100 experiments (different coefficients and noise). As expected, independent Lasso estimators do not benefit

**Figure 1**: 3 sets of color-labeled regression coefficients on a 2D grid. Each circle represents a non-zero coefficient. Different radii are used for a better distinction of overlapping features. (a) Inputs. Joint estimation of 3 ill-posed regression tasks using: (b) Lasso (c) MTW, latent Wasserstein barycenter shown in (d).
from task relatedness. Yet, they perform better than Dirty and MLL when supports poorly overlap, which confirms the results of Negahban and Wainwright [2008]. MTW however clearly wins in all scenarios.

4.2 Handwritten digits recognition.

We use the dataset of van Breukelen et al. [1998] consisting of handwritten numerals (‘0’–‘9’) extracted from a collection of Dutch utility maps. 200 patterns per class (for a total of 2,000 patterns) are digitized in binary images. We select 6 tasks corresponding to the digits (‘0’–‘5’) and the features corresponding to the pixel averages of (2 × 3) windows of the original (unprovided) (30 × 48) handwritten digit images, thus $p = 240$. We set $n = 10; 15; 20$ or 50 training samples per task. Model selection is carried out using a 5-folds cross-validation. We report in figure 3 the mean misclassification rate on the left-out validation set containing $n_v = 200 - n$ samples per task for 50 different random splits of the data. MTW is particularly efficient in the small $n$ regime with a significant 95% confidence interval. The regression coefficients obtained by each model are displayed in the appendix.

4.3 MEG source localization.

We use the publicly available dataset DS117 of Wake-man and Henson [2015]. DS117 contains MEG and EEG recordings of 16 subjects who underwent the same cognitive visual stimulus consisting in pictures of: famous people; scrambled faces; unfamiliar faces. Using the provided MRI scans, we compute the design matrices $X^t$ i.e. the forward operators of the magnetic field generated by a cortical triangulation of $p = 2101$ locations using the MNE software [Gramfort et al., 2013]. The regression outputs $Y^t$ correspond to measurements of the magnetic field on the surface of the scalp recorded by $n = 204$ sensors (we keep only MEG gradiometers), as for example used in [Owen et al., 2009]. Since the true brain activations $\theta^*$ are unknown, we quantify the performance of our model using the real $(X^t)_{t=1,…,T}$ and simulated $(Y^t)_{t=1,…,T}$. Note that the assumption of partial overlap is particularly adapted to this application. Indeed, while functional organization of the brain is comparable between subjects at a certain scale, one cannot assume that the activation foci are perfectly overlapping between individuals. In other words, active brain regions tend to be close in the population but not identical [Thirion et al., 2007, Xu et al., 2009].

Simulated activations. The regression coefficients (sources) are $k$-sparse ($k \in \mathbb{N}$), i.e all zero except for $k$ random locations chosen respectively in one of 11 distinct brain regions (displayed in supplementary material). Their amplitudes are taken uniformly within $20 – 30\text{nAm}$. Their sign is then decided by a coin toss (Bernoulli with 0.5 parameter). We generate in this manner a set of different regression coefficients for the number of tasks desired. We construct the outcome $Y^t$ with a SNR equal to 4. For MTW, the ground metric $M$ is the distance on the cortical mesh of $p$ vertices. It corresponds to the geodesic distance on the complex topology of the cortex.

Illustrative example. MTW is expected to be most valuable for non-overlapping supports. To visually illustrate the benefits of our model, we randomly select 2 subjects and simulate regression coefficients with 3 sources per task with only one common feature. Figure 4 shows MTW at its best: MTW leverages the geometrical proximity of the sources and thereby perfectly recovers the true supports. The independent Lasso estimator however reaches a poor AUC of 0.54. It selects features very far from the true brain regions which can lead to erroneous conclusions. Moreover, the
Wasserstein regularization for sparse multi-task regression

Figure 4: Each color corresponds to one of the two subjects (except for (d)). (a): True sources: one common feature (right side of the displayed hemisphere) and two non-overlapping sources. (b, c): Sources estimated by (b) Lasso and (c) MTW with the highest AUC score. (d) Shows the barycenter \( \bar{\theta} \) associated with MTW model. In this figure, the displayed activations were smoothed for the sake of visibility.

latent barycenter \( \bar{\theta} \) highlights the most representative sources of the cohort of subjects studied (Fig. 4 (d)).

Effect of degree of overlap. Using 3 subjects, we perform 30 trials with different noise and coefficients locations and values (Figure 5). We make the localization even harder by selecting 5 sources, i.e. 5 non-zero features per task. We select the best performance of all models in terms of AUC score. MTW outperforms all benchmarks in recovering the true supports in all scenarios. Unlike Dirty models, MLL fails to recover perfectly overlapping supports and has a large variance. This behavior may be due to the non-convexity of the penalty in (14) and potentially bad local minima.

Effect of number of tasks. When the number of non-zero features increases, recovering the support is more difficult. Figure 6 shows that MTW handles particularly well that scenario, as tasks increase. We compute the mean AUC score of 20 trials for 2, 4 and 8 tasks, 2 to 6 non-zero coefficients with an overlap of supports set to 50% and a SNR equal to 4. The curves obtained by independent Lasso overlap as it does not benefit from additional tasks. Dirty models handle relatedness through the \( \ell_1/\ell_2 \) penalty which only improves the estimation of the common features across tasks. This explains why the performance of Dirty models with 4 and 8 tasks is the same. MTW is unique in that it benefits from all 8 tasks.

Figure 5: Comparaison of different values of supports overlap using 3 tasks. Mean AUC score obtained on MEG data simulation with a SNR = 4 over 30 different experiments. MTW: Multi-task Wasserstein. MLL: Multi-level Lasso. MTW outperforms other models for all supports overlap fractions.

Figure 6: Mean AUC score for different numbers of tasks and support sizes with an overlap of 50% (20 different runs). MTW benefits more from additional tasks; obtained on MEG data simulation. SNR = 4.

Conclusion

The seminal work of Caruana [1993] has motivated a series of contributions leveraging the presence of multiple and related learning tasks (MTL) to improve statistical performance. Our work is one of them in the context of sparse high dimensional regression tasks where regressors can be associated to a geometric space. Using Optimal Transport to model proximity between coefficients, we proposed a convex formulation of MTL that does not require any overlap between the supports, contrarily to previous literature. We show how our Multi-task Wasserstein (MTW) model can be solved efficiently relying on proximal coordinate descent and Sinkhorn’s algorithm. Our experiments on synthetic and real data demonstrate that regardless of overlap, MTW leverages the geometry of the problem to outperform standard multi-task regression models.
Acknowledgments

MC and HJ acknowledge the support of a chaire d’excellence de l’IDEX Paris Saclay. AG and HJ were supported by the European Research Council Starting Grant SLAB ERC-YStG-676943.

References


M. A. Schmitz, M. Heitz, N. Bonneel, F. Ngolè, and D. Coeurjolly. Wasserstein Dictionary Learning: Op-
Wasserstein regularization for sparse multi-task regression


